2-25-2014

How Do Researchers Define Their Data Lifecycle and What Can We Learn from Their Definitions?

Jake R. Carlson
Purdue University, jakecar@umich.edu

How do researchers define their data lifecycle and what can we learn from their definitions?

Jake Carlson – Associate Professor of Library Science - Purdue University Libraries - jakecarlson@purdue.edu

PURDUE UNIVERSITY LIBRARIES
Access. Knowledge. Success.

INSTITUTE of Museum and Library SERVICES

Data Curation Profiles Directory

Data Curation Profile (DCP) capture information about a data set from the researcher's perspective. DCPs include a table of the stages of the data lifecycle as described by the researcher. DCPs also list which components of the data the researcher would be willing share with others (as shown by the grey shading).

The data tables from 32 DCPs are presented here. Comparing the content of these tables raises some important issues for service providers.

http://docs.lib.purdue.edu/dcp/

Agriculture / Land Use

Agricultural and Biological Engineering / Eco-Hydrology	Agronomy / Biofuels	Agronomy / Grain Yield	Agronomy / Land Use	Agronomy / Soil Micro Biology	Botany / Plant Taxonomy	Environmental Science / Herbivory	Plant Genomics	Plant Genetics / Corn	Plant Nutrition and Growth	Soil Ecology	Water Flow and Quality
Purdue	*Purdue*	*Purdue*	*Purdue*	*Purdue*	*Hawaii*	*Cornell*	*Purdue*	*Cornell*	*Purdue*	*Illinois*	*Purdue*
Data Collections and Calculations	Raw	Inheriting existing data set	Harvest	Raw	Primary Data	Raw - Field	Raw	Classification data	Raw	Raw 1	Raw
Data Collections and Calculations from Others	Trans position	Review and Repeat	Lab Work	Preparation / Compilation	Raw	Raw - Lab	Gathering Descriptive Information	Response Variables	Processed	Raw 2	Processed
Synthesis	Calculations & Conversions	Harvesting & Processing	Statistical Analysis	Statistical Output	Processed	Analyzed - Lab	Processing	Analysis	Integrated	Initial Digital Matrix	Interpolation
Parameterization	Statistical Analysis & Graphs	Statistical Analysis	Publication	Publication	Analyzed	Analyzed - Lab & Field	Ingest	Finalized	Extraction	Combined	Joined
Model Calibration Validation	Publications & Presentations	Aggregate			Finalized 1	Finalized	Internal Release for Analysis		Analysis	Corrected	Analyzed
Publication					Finalized 2		Public Release for Analysis		Qualitative	Output from SAS Software	Summarized
											Published

Earth Science

Carbonate Sedimentology	Geophysics and Semiology
Illinois	*Michigan*
Raw, hand collected data	Sample Collection & Indexing
Raw, sensor data	Tracking the sample and collecting analytical data
Raw, or "replicate" data	Modeling and interpretation
Reduced	Publication and dissemination
Field and micro-scope photos	
Digital Back-ups of Data - MS Word Files	
Digital Back-ups of Data - Digital Photos of the Field Notebook Pages	

Social Sciences & Humanities

Sociology / Demographics	Linguistics	Linguistics / Etymology	History / Sustainable Development	Architectural History / Epigraphy
Cornell	*Cornell*	*South Florida*	*Purdue*	*Tennessee*
Acquisition	Raw	Raw	Raw	Collecting Raw Inscription Data
Append	Processed	Processed	Processed	Encoding Inscriptions in TEI/XML
Aggregation	Analyzed	Analyzed	Analyzed	Final Published Data
Mapping	Finalized	Finalized	Finalized	
Estimations / Projections			History / Sustainable Development	

Genomics

Human Genomics	Human Cell Defense Systems	Movement of Proteins
Purdue	*Purdue*	*Purdue*
Reference	Explanatory	Raw
Raw	Record Keeping	Processed 1
Cleaned	Processed	Processed 2
Processed	Analyzed	Analyzed
Analyzed	Summarized	Published
	Published	

Engineering

Structural Control	Traffic Flow	Aerospace Engineering / Chemical Kinetics
Purdue	*Purdue*	*Michigan*
Data Acquisition	Raw	Raw: Pressure Trace
Conversion	Processing Stage 1	Raw: Concentrations of constituent reactants
Cleansing and Filtering	Processed	Raw: Video of Ignition Delay
Analysis & Plotting	Analyzed	Analyzed: Pressure Trace Data 1
Device Modeling & Data Comparison	Published	Analyzed: Pressure Trace Data 2
		Analyzed: Chromatographs
		3 additional stages - finalized

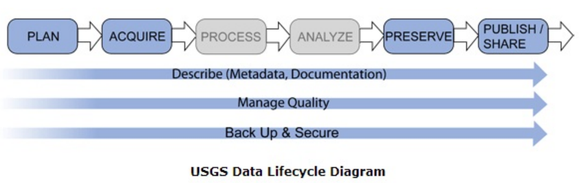
Life Sciences

Bio-Physics	Biochemistry / Histones	Bio-Mechanics Motion Studies
Cornell	*Purdue*	*Illinois*
Raw Data Acquisition	Methodology Development Part 1: Discovery	Raw
First pass analysis	Methodology Development Part 2: Refinement	Raw Filtered
Detailed Processing	Data Collection	Processed, Aggregate Data
Building Models	First Stage Data Analysis	Reduced
Fitting the Data to Models	Late Stage Data Analysis	Analytical Product
Validation of the Model / Publication	Publication	

Astronomy

Astronomy / Galactic Structure	Atmospheric Modeling	Astrophysics
RPI	*Illinois*	*UCSD*
1. Telescope Observations	Raw	Simulation
2a. Data analysis pipelines	Interpolated	Data Reduction
2b. Operational database	Intermediate	Analysis
3a. Calibration pipelines	Analyzed / Statistics	Publication
3b. Science use database (Analyzed and calibrated data)	Overview	
4a. SQL queries	Final	
5 additional stages		

Findings:



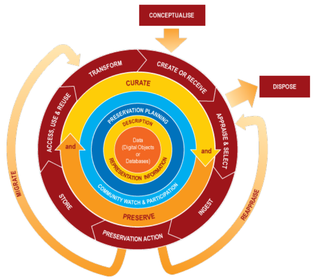
USGS Data Lifecycle Diagram

Researcher descriptions do not line up with data lifecycle models

Image from: http://www.usgs.gov/datamanagement/why-dm/lifecycleoverview.php

There is little alignment in what parts of the data to share

Image "Stick Figures Shake Hands" from: MS Office Clip Art



Curation is not included as a part of the data lifecycle

Image from: http://www.dcc.ac.uk/resources/curation-lifecycle-model

Direct communication with data producers is key

Image from: http://www.keleris.com